14D009 - Social and Economic Networks

Topic

Summary of random walk based centrality measures for complex networks

Author: Felix Gutmann
Student number: 125604
Program: M.S. Data Science
E-Mail: felix.gutmann@barcelonagse.eu

I Table of Contents

1 Introduction and review of relevant papers

A wide range of centrality measures have been proposed to find important nodes in a network. However, the term importance depends on the question one might to answer. The most natural way to think about importance is to count how many connections a node has. Besides that, two very prominent measures were proposed taking the global structure of the network into account.

Closeness centrality indicates how close a node is on average to all other nodes in the network. Therefore, it measures how quickly a node interacts with the rest of the network. [Wasserman and Faust, 1994, page 184].

On the other hand *(shortest path) betweenness centrality* counts how often a node lies on the shortest path among nodes in the network [Wasserman and Faust, 1994, page 190]. Hence, it can be interpreted as the power a node has on the flow of information in a network [Newman, 2005, page 2].

Both of the latter concepts assume that information is passed along the shortest path between two nodes and hence they are suitable to answer question where information transmission happens in such a way. However, that assumption might not always be valid. E.g. People discuss different topics with different friends and thus information spread could have a different nature. Following that, one might take also non geodesic paths into account. Another point to consider is that connection might change or disappear. Erasing a node from a network (e.g a friendship ends) can change the outcome for the mentioned centrality measures. Random walk based centralities provide a more robust centrality indicator [Mavroforakis et al., 2016, page 1] by taking all walks into account. This report provides a summary of some relevant concepts in that field.

The report has the following structure. First there will be an informal review of the papers and a related summary of key findings. Section two provides a formal summary of each concept. [1] Four papers have been reviewed.

To my knowledge [Noh and Rieger, 2004] and [Newman, 2005] can be viewed as the two "founding" papers in that field. The two concepts are inspired by the two concepts mentioned in the beginning. [Mavroforakis et al., 2016] is developing a random walk concept based on absorbing nodes. While the later three defining ran-

[1] Note that is aiming more on identifying the important steps for computations.

dom walks in the notion of *markow chains*, [Takes and Kosters, 2011] approaches random walks in a different way.

[Noh and Rieger, 2004] derive a centrality measure with special attention to *scale-free* networks.[2] Apart from that another motivation for their approach is, that employing "classical" centrality measures (such as closeness centrality) requires the global graph connectivity structure to be known in order to compute them. However given large networks such as the web graph or large social networks like facebook that obviously might be infeasible. As already mentioned in the introduction information might also not be transmitted over shortest paths. Therefore, information is considered moving randomly over all paths. The measure can be interpreted as how efficient nodes are in terms of receiving information from the network. Thus, it can be seen as variation of closeness centrality.
They study the theoretical behavior of the centrality measure by simulation (Barabasi-Albert networks).[3] They find that the random walk centrality is mostly determined by the degree distribution of the network. The centrality derived in the paper is restricted to undirected networks.
To my knowledge they don't provide a implementation. Furthermore, a discussion of computational complexity is missing.[4]

[Newman, 2005] provides a measure for betweenness centrality. It is inspired both by shortest path and flow betweenness centrality.[5]. Despite the fact that flow betweenness centrality also takes nodes into account, which do not lie on the shortest path, both concepts kind of assume an optimal movement of information through the network.
Likewise random walk centrality their concept considers information diffusing randomly over the network and therefore it also considers walks apart from shortest paths. Essentially it indicates how often a vertex will lie on average on a random walk between all pairs of source and target nodes. However, shortest path are still

[2] The degree distribution follows a power law distribution

[3] The results require some more theoretical knowledge and therefore are not shown at this place. They can be found in [Noh and Rieger, 2004, page 4]

[4] Since the computation **in matrix notation** only requires inverting a matrix (see equation (5)) the complexity can be assumed to be somewhat like $\mathcal{O}(n^3)$.

[5] The latter considers the flow through k in the maximal flow from node i to j as an indicator of centrality

contribute more to centrality.
Besides a formal derivation the paper studies some properties. The paper identifies (especially) high correlation with shortest path betweenness centrality. In that context an interesting example is provided to justify the application. They study a network of the spread of sexual diseases. It is revealed that the random walk measures identifies additionally nodes, which are not recovered by shortest path betweenness centrality.[6]

[Mavroforakis et al., 2016] has a more theoretical objective. The paper aims to provide a random walk based centrality concept, which can be e.g. applied for searching algorithms. The challenge is to find a set of nodes, which matches a users search and find the most central ones among those nodes. They derive an objective function to find such a set k-central nodes (the objective function is presented later on). In further context the paper lines out important properties, proofs and algorithms. In particular it provides a greedy algorithm to solve the optimization problem. Despite the fact the problem turns out to be NP-hard, two key properties of the objective function (monotonicity and super modularity) ensure an approximation guarantee. Since the algorithm requires costly matrix inversions they show how to employ Sherman Morrison inversion technique to speed up computation. Finally they study several heuristic techniques, which can additionally speed up computation. By studying performance on several available networks they find that most of the heuristic track the results from the greedy solution (for example*Personalized Pagerank*, *Degree and distance centrality*.[7]

While the last three papers approach the issue with techniques related to markov chains, [Takes and Kosters, 2011] interpret random walks in the following way. In general their algorithm runs for a fixed time and samples in each iteration the next node to be chosen. This probability of the next node is determined by a weighted combination of *degree centrality* and *neighborhood density* (see equations (12), (13) and (14)). Furthermore, an additional random element extend this concept to ex-

[6] Other examples are an application to "Florentine Families" network and "co-authorship" network, which are not that interesting

[7] The authors provide a python implementation

plore the graph. Their method is studied with a subset of the dutch social network ”HYVES” with 8 million people. They define a set of 4, 867 (0.06%) people to be prominent (politician, artists etc.). Considering that as the ground truth they let the algorithm compete against various other measures (Degree Centrality, Random Walk, PageRank, HITS). Their algorithm outperforms all of the latter beating the best (Degree Centrality) significantly with different variations of that data set and based on recall, precision or F-measure as evaluation measures.

2 Mathematical background of centrality measures

This section provides the formal derivations for each centrality measure. As a prerequisite notations for graphs are introduced. Furthermore, some background knowledge on markov chains is provided to give some intuition behind the used terminology later on.[8]

2.1 Notation and basic background of markov chains

First we might formally introduce a graph. A graph is an ordered pair $G = (V, E)$, where V is the set of nodes with $V = \{v_1, \dots\}$ and E is the set of edges with $E = \{e_1, \dots, e_n\}$. The adjacency matrix of such a graph G is denoted as $\mathbf{A}$, where entries a_{ij} are defined as [Aigner, 2007, page 124]:

$$a_{ij} = \begin{cases} 1 & \text{if } \mathrm{v}_i\mathrm{v}_j \in \mathrm{E} \\ 0 & \text{otherwise} \end{cases}$$

The *neighborhood* of a node $N(v_i)$ is the set of vertices adjacent to v_i, so precisely the set $N(v) = \{u \in V : (u, v) \in E\}$. Following that, the *degree* of a node is defined as:

$$d(v_i) = |N(v_i)| \tag{1}$$

[8] Note that notation sometimes deviates from the original papers to set up consistency.

Using equation (1) we define **d** as the n × 1 vector of all degrees, $\bar{d} = \sum_{i=1}^{N}$ as the sum of degrees and **D** as the diagonal matrix of the degrees. Finally, denote N as the number of nodes defined by $N = |V|$.

A random walk is characterized as follows. It starts at a given node of the graph and moves to an adjacent node with a certain probability. Such random walks are usually modeled with the notion of markov chains. All following definitions can be found in chapter 23.2 in [Wasserman, 2004, page 383 et. seqq.]. Consider the set of N possible different states $\mathcal{X} = \{1, \dots, N\}$.
A markov chain is modeled with an N × N *transition matrix* **P**, where an entry denotes the probability to move from a given state to another. It said to be *homogeneous*, if transition probabilities are not changing over time.[9] A random walks is a stochastic process and therefore evolves over a number of (in our case) discrete index steps denoted by $T = \{1, 2, \dots\}$. Combining that, an entry of the transition matrix satisfies:

$$p_{ij} = \mathbb{P}\left(X_{t+1} = j | X_t = i\right)$$

The next state is modeled considering only the last state. A state is said to be *absorbing* if an entry of the transition matrix is equal to one.
We want to model the behavior over time. The transition probabilities in the t-th step is obtained by simple matrix multiplication of transition matrix:

$$\mathbf{P}(t) := \underbrace{\mathbf{P} \times \cdots \times \mathbf{P}}_{\text{t - times}} \tag{2}$$

A state is called *recurrent* if in some step for $t \geq 1$ there is a probability that the random walks returns to its initial state. A recurrent state therefore satisfies:

$$\mathbb{P}\left(X_t = i | X_0 = i\right) = 1$$

If state is **not** recurrent it is called *transient.* In that context we might ask the question how much time it takes until a random walk returns to its initial state. Hence, we introduce *recurrence time* and the *mean recurrence time.* The recurrence

[9] In the following this will be the case

time is define as

$$T_{ij} = \min\{t > 0 : X_t = j\}$$

For a given step in time t and a recurrent state we can compute the expected value of the recurrence time as:

$$\mathbb{E}(\mathrm{T}_{ii}) = m_i = \sum_{\mathrm{t}} \mathrm{t} f_{ii}(\mathrm{t}) \tag{3}$$

where

$$f_{ij}(\mathrm{t}) = \mathbb{P}\left(X_1 \neq j, X_2 \neq j, \ldots, X_t = j | X_0 = i\right)$$

Finally we want to give guidance how to simulate a markov chain. Therefore, we introduce the marginal probability of a state.

$$\mu_0 = \mathbb{P}\left(X_0 = i\right)$$

The simulation follows a simple procedure. Let $\boldsymbol{\mu}_0$ the vector of marginal probabilities at $t = 0$, such that $\boldsymbol{\mu}_0 = (\mu_{0,1}, \ldots, \mu_{0,n})$. To simulate the probabilities for the given states after t steps we just need to multiply the marginal probabilities with the transition matrix after t steps and so:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 \mathbf{P}(t)$$

2.2 Random walk centrality

The paper especially introduces an explicit expression of the *Mean first passage time* to model their centrality centrality (the idea seem to be related to (3)).[10] The graph is assumed to be connected, and if not so the procedure is conducted for each component. We start by defining the transition probabilities for the random walk by normalizing the the entries of the adjacency matrix by the degree.

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A} \tag{4}$$

Next we define the normalized degree of a node as follow :

$$\mathbf{P}_i^{\infty} = \frac{d_i}{\sum_{i=1}^{N}} = \frac{d_i}{\bar{d}}$$

It can be shown that this is the stationary distribution of the process. The authors claim that random walks on finite networks are recurrent and the Mean First Passage Time is:

$$T_{ij} = \sum_{t=0}^{\infty} \mathrm{t} F_{ij}(t)$$

Following that one can show that the mean first passage time in this case can be explicitly expressed in the following way (see in particular [Noh and Rieger, 2004, page 2]):

$$T_{ij} = \begin{cases} \frac{\bar{d}}{d_i} & \text{for j = i} \\ \frac{\bar{d}}{d_i}\left[R_{jj}^{(0)} - R_{ij}^{(0)}\right] & \text{for j} \neq \text{i} \end{cases}$$

[10] This concept and the following one got for example adapted and adjusted to compute central sectors in Input Output Tables by [Blöchl et al., 2010] and [Blöchl et al., 2011])

T_{ii} is the average return time, which only depends on the degree of a node and $R_{ij}^{(0)} = \sum_{t=0}^{\infty} (p_{ii}(t) - \mathbf{P}_i^{\infty})$. The mean first passage time is not symmetric. One can show the following identity:

$$T_{ij} - T_{ji} = \bar{d}\left(\frac{R_{jj}^{(0)}}{d_j} - \frac{R_{ii}^{(0)}}{d_i}\right) - \bar{d}\left(\frac{R_{ij}^{(0)}}{d_j} - \frac{R_{ji}^{(0)}}{d_i}\right)$$
$$T_{ij} - T_{ji} = C_j^{-1} - C_i^{-1}$$

Where C is the centrality of a node. We see from the last equation the difference in mean first passage time is low for high centrality values (freely spoken) and therefore the speed of information transmission is determined by the last equation. The final measure C_i is then defined as:

$$C_i = \frac{\mathbf{P}_i^{\infty}}{R_{ii}^{(0)}} \tag{5}$$

As stated in the summary, the centrality expresses how fast a node receives information from the network Higher values in (5) are therefore indicating higher centrality.

2.3 Random walk betweness centrality

In the paper the centrality concept is derived with an analogy of an electrical flow network. The authors show that this is equivalent to the random walk interpretation. In the following this part is skipped and only a short guidance for computation is provided (furthermore, the intuition of that approach is not really obvious). Again the graph is assumed to be connected and in case it is not, the procedure should be repeated for each component. Start by removing arbitrarily a row and corresponding column denoted as t from the transition matrix. Hence we denote this matrix without those entries as:

$$\mathbf{P}_t = \mathbf{A}_t \mathbf{D}_t^{-1} \tag{6}$$

We are interested in how often a node is passed on random walks averaged over all source target pairs of nodes. In matrix terms we can express this in the following way:

$$\mathbf{V} = \mathbf{D}_t^{-1} \left(\mathbf{I} - \mathbf{P}_t\right)^{-1} \mathbf{s} = \left(\mathbf{D}_t - \mathbf{A}_t\right)^{-1} \mathbf{s}$$

The matrix $\mathbf{V}$ is the voltage matrix and s is defined in the following way.[11]

$$\mathbf{s} = \begin{cases} +1 & \text{if } i = s \\ -1 & \text{if } i = t \\ 0 & \text{otherwise} \end{cases}$$

In the next step add a row and a column of zeros back into the matrix $\left(\mathbf{D}_t - \mathbf{A}_t\right)^{-1}$ at position t and we denote this as $\mathbf{T}$. As a last step we introduce I to solve for betweenness.

[11] The name comes from the electrical network analogy

$$I_i^{(st)} = \frac{1}{2} \sum_j A_{ij} |V_i^{(st)} - V_j^{(st)}| = \frac{1}{2} \sum_j A_{ij} |T_{is} - T_{it} - T_{js} + T_{jt}|$$

and we set $I_s^{(st)} = I_t^{(st)} = 1$. Finally, we compute the betweenness of a node as:

$$b_i = \frac{\sum_{s<t} I_i^{(st)}}{1/2n(n-1)} \tag{7}$$

2.4 Absorbing random walk centrality

For this measure we may have to introduce some additional notation. The objective is to to identify a set of k central node. Denote this set of central nodes as C, where c $\in$ C and C $\subseteq$ V.
Moreover, define the *query nodes* Q such that $Q \subseteq V$.
We define a third set of so called *candidate nodes*, which are potentially in C, such that $C \subseteq D$. This distinction serves the purpose that depending on the application the set of central nodes can be limited to the nodes in Q, but in others can be potentially belong to whole V.
The random walk in this concept starts at a node in the query nodes $q \in Q$. The random walks proceeds as long as it arrives at any node in C, where it gets absorbed. The starting node is chosen with a discrete distribution $s(v_i)$.[12]. Likewise the other centralities the centrality here is defined as the expected value how fast that happens. Since a node in C absorbs the random walk the probability of escaping is zero and therefore entries $p_{cj} = 0$ and $p_{cc} = 1$. Non absorbing nodes are considered to be transient nodes and denoted as T = V $\setminus$ C.
An extension to the algorithm is that a random walk can be restarted with a given fixed probability α. Taking the last mentioned facts we write out the adjusted transition probabilities for

$$p_{ij} = \begin{cases} \alpha s(v_j) & \text{if } v_j \in Q \setminus N(v_i) \\ \frac{(1-\alpha)}{d_i} + \alpha s(v_j) & \text{if } v_j \in N(v_i) \end{cases} \tag{8}$$

Finally the complete transition matrix can be expressed as a block wise arrangement of the following four sub matrices:

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{\mathrm{TT}} & \mathbf{P}_{\mathrm{TC}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{9}$$

[12] In the simplest case this can be taken as a uniform distribution

The probability that the random work after t has't been absorbed is $\mathbf{P}_{\mathrm{TT}}(t)$. After infinite steps this is computed with the geometric series:

$$\mathbf{F} = \sum_{t=0}^{\infty} \mathbf{P}_{\mathrm{TT}}(t) = (\mathbf{I} - \mathbf{P}_{\mathrm{TT}})^{-1} \tag{10}$$

Using the previous equation (10) that the expected length of the random walk getting absorbed can be computed using the following vector:

$$\mathbf{L} = \mathbf{L}_C = \begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix} \mathbf{1}$$

The final measure is achieved by summing over all nodes in Q.

$$C_{RWA} = \mathbf{s}^{\mathrm{T}} \mathbf{L}_C \tag{11}$$

However, this measure comes with a cost. Not only follows from equation (10) that each computation has to be done inverting a matrix, which can be computationally expensive, but also we have to optimize this procedure over different choices of C.

2.5 Biased random walk centrality

As stated in section 1. this paper has a slightly different approach. Algorithm one gives a pseudo code for the biased random walk centrality. The algorithm requires as input an unweighted Graph - G, the number of "random" steps done - N, a weighting parameter - α (exact meaning see later on) and a probability bound - p. In the first step assign zero as centrality to all nodes. Then run the algorithm for N steps (where N should be much larger than number of nodes). Per iteration update the value of the current node. Then sample a uniform number between zero and one. If this number is higher than the threshold pick a node randomly of the set of nodes. If not pick the next node according to the BiasSelectFrom() function which selects the next node according to the probability computed in (14). After N steps the functions returns a vector with prominence values for each node.

```
input  : G,N,$\alpha$,p
output: Importance values for each node $v$
for $v \in V$ do
  | f(v) = 0
end
i = 0;
v = RandomNodeFrom(V) ;
while $i < N$ do
  | $f(v) \leftarrow f(v) + \frac{1}{N}$ ;
  | if rand(0,1)<$p$ then
  |   | v ← BiasSelectFrom(N(v),$\alpha$)
  | else
  |   | v ← RandomSelectFrom(V)
  | end
  | i ← i + 1
end
return f(v)
```

Algorithm 1: Guided Random Walk

For defining the BiasSelectFrom() function we may first introduce two necessary concepts. First we define the importance of a node based on the number of connections. This stems from the fact that the number of connections might be a important prominence indicator in the context of social networks.

$$f_{deg}(v) = 1 - \frac{1}{|N(v)|} \tag{12}$$

However, this concept can be extended. The authors define *neighberhood density* as:

$$\begin{aligned} f_{nd}(v) &= 1 - \sum_{w \in N(v)} \frac{|N(w) \cap N(v)|}{(|N(w)| - 1)|N(v)|} \\ &= 1 - \sum_{w \in N(v)} \frac{|N(w) \cap N(v)|}{(\mathrm{d}_w - 1)\mathrm{d}_v} \end{aligned} \tag{13}$$

Finally the last two equations are used to compute the probability for the BiasSelectFrom() function. The parameter α controls the how much should be given to the prominence concepts in equation (12) and (13). Setting α close to one will result in values close to degree centrality. Hence we compute the probability of a node $w \in N(v)$ to be chosen as follows, which is the basis to execute the BiasSelecFrom().

$$\mathbb{P}(w) = \frac{\alpha f_{deg}(w) + (1 - \alpha) f_{nd}(w)}{\sum_{u \in N(v)} (\alpha f_{deg}(u) + (1 - \alpha) f_{nd}(u))} \tag{14}$$

3 List of Literature

[Aigner, 2007] Aigner, M. (2007). *Discrete Mathematics.* American Mathematical Society, Providence.

[Blöchl et al., 2010] Blöchl, F., Fisher, E. O., and Theis, F. (2010). Which Sectors of a Modern Economy are most Central? *CESifo Working Paper Series*, (3175).

[Blöchl et al., 2011] Blöchl, F., Theis, F. J., Vega-Redondo, F., and Fisher, E. O. (2011). Vertex Centralities in Input-Output Networks, Reveal the Structure of Modern Economies. *Phys. Rev. E*, 83(4).

[Mavroforakis et al., 2016] Mavroforakis, C., Mathioudakis, M., and Gionis, A. (2016). Absorbing random-walk centrality: Theory and algorithms. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2016-Janua:901–906.

[Newman, 2005] Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54.

[Noh and Rieger, 2004] Noh, J. D. and Rieger, H. (2004). Random Walks on Complex Networks. *Physical Review Letters*, 92(11):118701–1.

[Takes and Kosters, 2011] Takes, F. W. and Kosters, W. A. (2011). Identifying prominent actors in online social networks using biased randomwalks. *Belgian/Netherlands Artificial Intelligence Conference.*

[Wasserman, 2004] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*, volume 1542.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, Cambridge.